A Multi-task Mean Teacher for Semi-supervised Facial Affective Behavior Analysis

Lingfeng Wang, Shisen Wang

School of Information and Communication Engineering, University of Electronic Science and Technology of China

Abstract

Affective Behavior Analysis is an important part in human-computer interaction. Existing successful affective behavior analysis method such as TSAV[9] suffer from challenge of incomplete labeled datasets. To boost its performance, this paper presents a multi-task mean teacher model for semi-supervised Affective Behavior Analysis to learn from missing labels and exploring the learning of multiple correlated task simultaneously. To be specific, we first utilize TSAV as baseline model to simultaneously recognize the three tasks. We have modified the preprocessing method of rendering mask to provide better semantics information. After that, we extended TSAV model to semi-supervised model using mean teacher, which allow it to be benefited from unlabeled data. Experimental results on validation datasets show that our method achieves better performance than TSAV model, which verifies that the proposed network can effectively learn additional unlabeled data to boost the affective behavior analysis performance.

I. INTRODUCTION

Facial affective behavior recognition plays an important role in human-computer interaction[1]. In this way Intelligent systems can benefit from the ability to understand human feelings and behaviors, which makes human computer interaction more applicable.

There are different methods of Representing human emotions such as valence/arousal value, action unit (AU), and facial expression. Valence represents how positive the person is while arousal describes how active the person is. AUs are the basic actions of individuals or groups of muscles for portraying emotions. As for facial expression, it classifies into seven categories, neutral, anger, disgust, fear, happiness, sadness, and surprise.

The challenges for ABAW ICCV-2021 Competition [1][2][3][4][5][6][7][8]include valence-arousal estimation, facial action unit detection, and expression classification. There are strong correlation between the three different tasks. Multi-task learning can learn to extract features from correlated tasks, and has been proven to provide better performance than training on a single task. Among methods in the last year's competition, Two-Stream Aural-Visual model (TSAV)[9] proposed achieved superior performance in a multi-task manner. However, most samples in Aff-Wild2 dataset are labelled for only one or two task. Only limited number of samples are labeled completely for all three tasks. That's to say, there are different number of labeled data for the three tasks. During the multi-task training process, labeled data could be enough for one task while be insufficient for other tasks, which leads to imbalanced performance among different tasks. The authors of TSAV faced this challenge and had to create additional pseudo labels for model training.

To tackle this problem, we develop a multi-task mean teacher[10] framework for boosting affective behavior recognition performance. We first adopt TSAV model as baseline model. for mutually learning three tasks. ,The usage of mask as input is believed to be most helpful to the performance[9]. To this regard, we use an improved method of rendering mask to provide better semantics information. Second, we take this multi-task model as both the student network and the teacher network. We then propose a supervised multi-task loss for labeled data to integrate the supervised losses on all three tasks. After that, we enforce the three tasks' results of the student network and the teacher network to be consistent, respectively, on all the unlabeled data. By adding the supervised loss and the consistency loss from the three tasks to train the model, our network can be trained from both labeled and unlabeled data.

Our major contributions are summarized as:

• First, we propose a method to enhance the performance of TSAV by using improved rendered mask as input.

• Second, instead of using the complex pseudo label described in TSAV, we design a multi-task mean teacher framework to fuse consistency loss of unlabeled data from three prediction tasks for shadow detection. In this way, multi-task model can benefit from both labeled and unlabeled data.

II. RELATED WORKS

In recent years, most of the existing research for human affect focused on valence-arousal estimation (VA), facial action unit (AU) detection, and facial expression (EX) classification. We will introduce the latest related work briefly.

Kossaif et al. [1] proposed a dataset for valence-arousal estimation called AFEW-VA and demonstrated the representational power of geometric features. Kollias et al.[2] extend the large-scale database(Aff-Wild) [3] to study continuous emotions called Aff-Wild2. Aff-Wild2 is the first ever database annotated for all three main behavior tasks: VA,

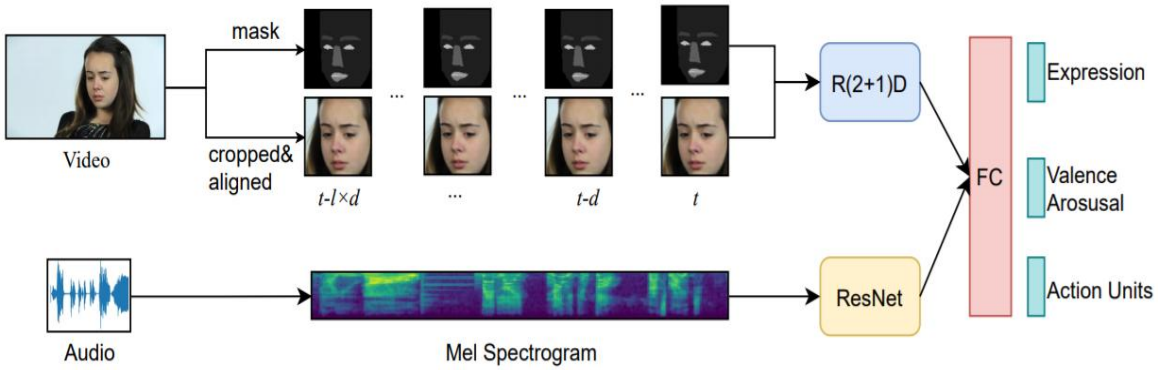

Fig. 1. Framework for multi-task affective behavior analysis model

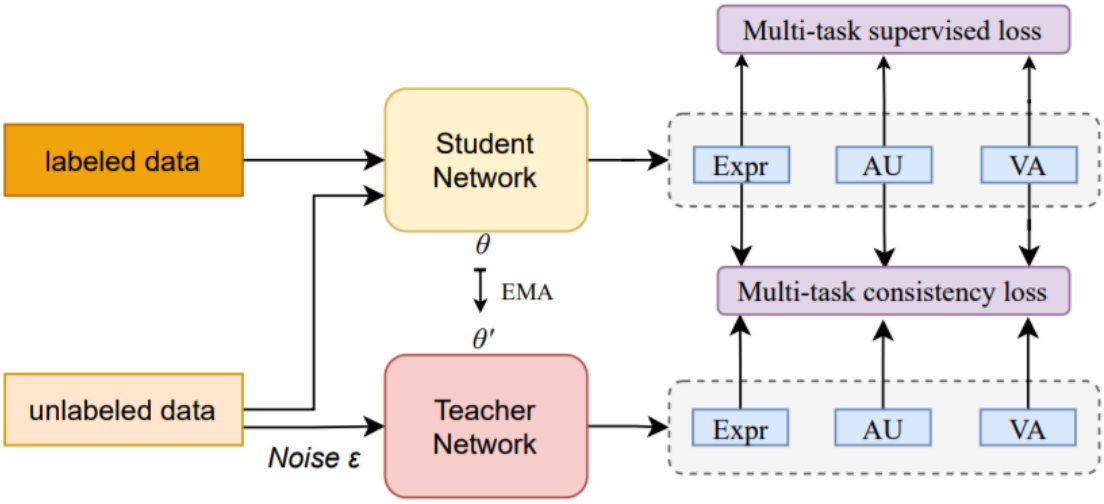

Fig. 2. Framework for Mean Teacher

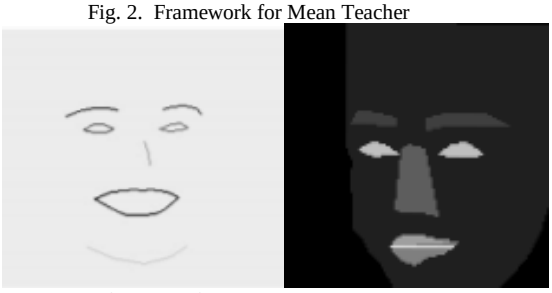

Fig. 3. Left: Mask of TSAV vs Right: Proposed mask

AU and EX. In [4], Kollias et al. proposed FaceBehaviorNet for large-scale face analysis, by jointly learning multiple facial affective behavior tasks and a distribution matching approach. Chang et al. [13] propose an integrated deep learning framework for facial attribute recognition, AU detection, and VA estimation by applying AU to estimate the VA intensity. Pan et al. [14]designed a framework to aggregate spatial and temporal convolutional features across the entire extent of a video. Kim et al. [15]introduced adversarial learning to solve facial emotion recognition problems, which enabled the model to better understand complex emotional elements inherent in strong emotions. In addition, they proposed a contrastive loss function to improve efficiency for adversarial learning. Li[16] use MIMAMO Net [17] to extract micro-motion and macro-motion information for improving Concordance Correlation Coefficient (CCC) for valence and arousal. Deng et al. [18] use a data-driven teacher model to fill in the missing labels.

III. METHODOLOGY

A. Multi-task Affective Behavior Recognition Model

Fig.1 shows the framework of our multi-task affective behavior analysis model. All the video clips in the competition dataset are splitting into image and audio streams. These streams are pre-processed individually and then synchronously fed into the aural-visual model. Finally, the model output joint prediction of three different emotion representations.

For the Visual stream, the input frames are cropped facial region images. These facial crops are all aligned according to 5 point template (eye centers, nose tip, outer mouth corners). Additionally, the usage of mask in TSAV is believed to be most helpful to its performance. To further enhance its performance, we use HRNet [11] to detect 106 facial landmarks for every face. With these landmarks, we can render a mask image of facial segmentation result. As is shown in Fig.3, comparing to the mask rendering method in TSAV, which can only render eye contours, the nose, the chin, the brows, and the outer lip contour

based on the 68 landmarks, our method can provide better semantics information.

In this way, each frame image has 4 channels (RGB + mask), while an input clip contains l frames. All the frames are sampled with dilation d. Here we choose clip length l = 8 and dilation d = 6.

As for audio stream, we compute a mel spectrogramm for all audio stream extracted from the video using TorchAudio package. For each clip, spectrogram is cut into a smaller sub-spectrogram with the center of sub-spectrogram aligning with the current frame at time t.

The two stream are input to TSAV model. TSAV employ (R2+1)D [12]model to extract spatio-temporal information from visual stream as well as resnet-18 for mel spectrogram analysis. Finally, the outputs of both sub-models are merged and give the joint prediction of three different expression representations (Continuous valence and arousal, basic expression and action units).

B. Mean Teacher

Mean teacher framework[10] is extended from supervised architecture by make a copy of original model. The original model is called student and the new one is called the teacher. The parameters of the teacher network in each training step, are updated via the exponential moving average (EMA) strategy. The parameters of the teacher network at the t training iteration are

$$\theta'_t = \eta\theta'_{t-1} + (1-\eta)\theta_t$$

θ_t represent parameter of model while η is hyper parameter of moving average. Here we choose $\eta = 0.99$.

At each training step, use the same minibatch as inputs to both the student and the teacher but add noise to the teacher model. Here we apply random brightness augmentation for each input clip of teacher model.

For the unlabeled data, we pass it into the student and teacher networks to obtain prediction for three tasks. We take these predictions as hard label and then enforce the predictions from the student network and teacher network to be consistent, resulting in a multi-task loss. Let the optimizer update the student weights normally.

After each training step, update the teacher weights a little bit toward the student weights by calculating the exponential moving average (EMA) of the student weights.

C. Loss Function

For such a multi-task learning model, each task have its loss respectively. For categorical expression classification task, we use categorical cross entropy. The binary cross entropy is used for action unit detection and the concordance correlation coefficient loss for valence and arousal estimation.

For all the labeled samples in the current mini-batch, we calculate supervised loss by adding the losses for expression, action unit, and valence and arousal estimation tasks.

$$L^s = L^s_{expr} + L^s_{au} + L^s_{va}$$

As for the unlabeled samples, we take the prediction results of teacher model as hard label, and calculate losses between hard label and prediction of student model in the same way.

$$L^c = L^c_{expr} + L^c_{au} + L^c_{va}$$

The sum of supervised loss for labeled samples and consistency loss for unlabeled samples is the final total loss for current batch.

$$L_{total} = L^s + L^c$$

IV. Experimental

A. Dataset

We only use the large-scale in-the-wild Aff-Wild2 dataset for our experiments. This dataset contains 564 videos with frame-level annotations for valence-arousal estimation, facial action unit detection, and expression classification tasks. We randomly split samples in each task into train set and validation set at ratio of 8:2

B. Model

Model is trained our train split dataset only. We use the pretrained weight from TSAV. We didn't adopt the data preprocessing step of filter and pseudo labels described in [9] to evaluate the semi-supervised performance of proposed method. Model is optimized using Adam optimizer and a learning rate of 0.0005.

C. Result

Table 1 shows results of TSAV and proposed method in validation dataset. Since the test dataset is not released, we trained TSAV and proposed method on our train split and evaluated their performance using our validation dataset. The performance of baseline is from [1]

TABLE I
PERFORMANCE OF MODELS ON VALIDATION SET

Model	Expression Criterion	CCC Mean	Action Unit Criterion
Baseline	0.36	0.22	0.31
TSAV	0.508	0.537	0.623
Proposed	0.517	0.562	0.782

The result indicates that our method significantly surpasses the baseline result and outperform TSAV especially in AU task. Our improved facial mask images most likely help the performance since it can provide a stronger prior for AU key points.

V. CONCLUSION

This paper presents a semi-supervised facial affective behavior recognition model by developing a multi-task mean teacher framework. Our key idea is to firstly use improved facial mask to provide a stronger prior and enhance performance of model. Then we employ the mean teacher semi-supervised learning to learn additional unlabeled data for further improving the recognition performance.

Experimental results on validation datasets show that our model outperforms original TSAV model in all task, especially AU classification, which verifies the effectiveness of proposed method.

REFERENCES

:

[1] *D. Kollias, et. al.: "Analysing Affective Behavior in the second ABAW2 Competition", 2021*

[2] *D. Kollias, et. al.: "Analysing Affective Behavior in the First ABAW 2020 Competition". IEEE FG, 2020*

[3] *D. Kollias, et. al.: "Distribution Matching for Heterogeneous Multi-Task Learning: a Large-scale Face Study", 2021*

[4] *D. Kollias,S. Zafeiriou: "Affect Analysis in-the-wild: Valence-Arousal, Expressions, Action Units and a Unified Framework, 2021*

[5] *D. Kollias, S. Zafeiriou: "Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace". BMVC, 2019*

[6] *D. Kollias, et at.: "Face Behavior a la carte: Expressions, Affect and Action Units in a Single Network", 2019*

[7] *D. Kollias, et. al.: "Deep Affect Prediction in-the-wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond". International Journal of Computer Vision (IJCV), 2019*

[8] *S. Zafeiriou, et. al. "Aff-Wild: Valence and Arousal in-the-wild Challenge", CVPRW, 2017*

[9] *F. Kuhnke, L. Rumberg and J. Ostermann, "Two-Stream Aural-Visual Affect Analysis in the Wild," in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG), Buenos Aires, undefined, AR, 2020 pp. 600-605.*

[10] *Tarvainen A , Valpola H . Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[J]. 2017.*

[11] *Cheng B , Xiao B , Wang J , et al. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.*

[12] *D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6450–6459, 2018.*

[13] *Chang W Y, Hsu S H, Chien J H. FATAUVA-Net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017: 17-25.*

[14] *Pan X, Ying G, Chen G, et al. A deep spatial and temporal aggregation framework for video-based facial expression recognition[J]. IEEE Access, 2019, 7: 48807-48815.*

[15] *Kim D H, Song B C. Contrastive Adversarial Learning for Person Independent Facial Emotion Recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(7): 5948-5956.*

[16] *Li I. Technical Report for Valence-Arousal Estimation on Affwild2 Dataset[J]. arXiv preprint arXiv:2105.01502, 2021.*

[17] *Deng D, Chen Z, Zhou Y, et al. Mimamo net: Integrating micro-and macro-motion for video emotion recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(03): 2621-2628.*

[18] *Deng D, Chen Z, Shi B E. Multitask emotion recognition with incomplete labels[C]//2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020: 592-599.*